

# CUNY SPS DATA 621 - CTG5 - HW1

*Betsy Rosalen, Gabrielle Bartomeo, Jeremy O'Brien, Lidiia Tronina, Rose Koh*

*February 27, 2019*

## Contents

<b>1</b>	<b>DATA EXPLORATION</b>	<b>1</b>
1.1	Instructions: (Betsy: we can take these out later, but it helps me to see them right here for now...)	1
1.2	Possible writeup for Data Exploration	2
1.3	Describe the size:	2
1.4	Statistics summary	2
1.5	Data Visualization	2
<b>2</b>	<b>DATA PREPARATION</b>	<b>7</b>
2.1	Instructions:	7
2.2	Possible writeup for Data Preparation	7
2.3	Information in general, don't use the writeups here as-is	8
2.4	Missing Values	8
2.5	NA Imputation	8
2.6	Remove Outliers	9
2.7	Feature Engineering	10
<b>3</b>	<b>BUILD MODELS</b>	<b>11</b>
3.1	Instructions:	11
3.2	MODEL 1	11
3.3	MODEL 2	16
3.4	MODEL 3	28
<b>4</b>	<b>SELECT MODELS</b>	<b>29</b>
4.1	Instructions:	29
4.2	Comparison of models	30
<b>5</b>	<b>Multi-collinearity</b>	<b>30</b>

## 1 DATA EXPLORATION

### 1.1 Instructions: (Betsy: we can take these out later, but it helps me to see them right here for now...)

Describe the size and the variables in the moneyball training data set. Consider that too much detail will cause a manager to lose interest while too little detail will make the manager consider that you aren't doing your job. Some suggestions are given below. Please do NOT treat this as a check list of things to do to complete the assignment. You should have your own thoughts on what to tell the boss. These are just ideas.

- a. Mean / Standard Deviation / Median
- b. Bar Chart or Box Plot of the data and/or Histograms
- c. Is the data correlated to the target variable (or to other variables?)
- d. Are any of the variables missing and need to be imputed 'fixed'?

## 1.2 Possible writeup for Data Exploration

Professionals and gamblers alike are always seeking to optimize their chances of winning, whether it be sports, games, or their bets on them. Major League Baseball is a multibillion dollar industry where individual teams, players, and those who profit off of their success stand to benefit most from such optimization.

In order to determine the best way to infer whether the 162 games in a baseball team's year will result in more wins overall, data from 1871 to 2006 where each set of values represented a season for an unnamed team, totalling 2,276 records. For each team their number of wins in a given year were given with a maximum possible of 162 wins, in addition to that team's base hits, doubles, triples, homeruns, walks, and strikeouts by batters, batters hit by pitches, bases stolen by batters and the number of times they were caught stealing, the number of errors, double plays, walks, hits, and homeruns allowed, and strikeouts by pitchers.

## 1.3 Describe the size:

The money ball data is 144kb in size. The data contains 2,276 rows and 16 columns without the index. The variables are continuous integer. The TARGET\_WINS is our response variable. There are 3,478 missing values out of 36,416 observations.

```
## [1] 2276   16  
## [1] 259   15
```

## 1.4 Statistics summary

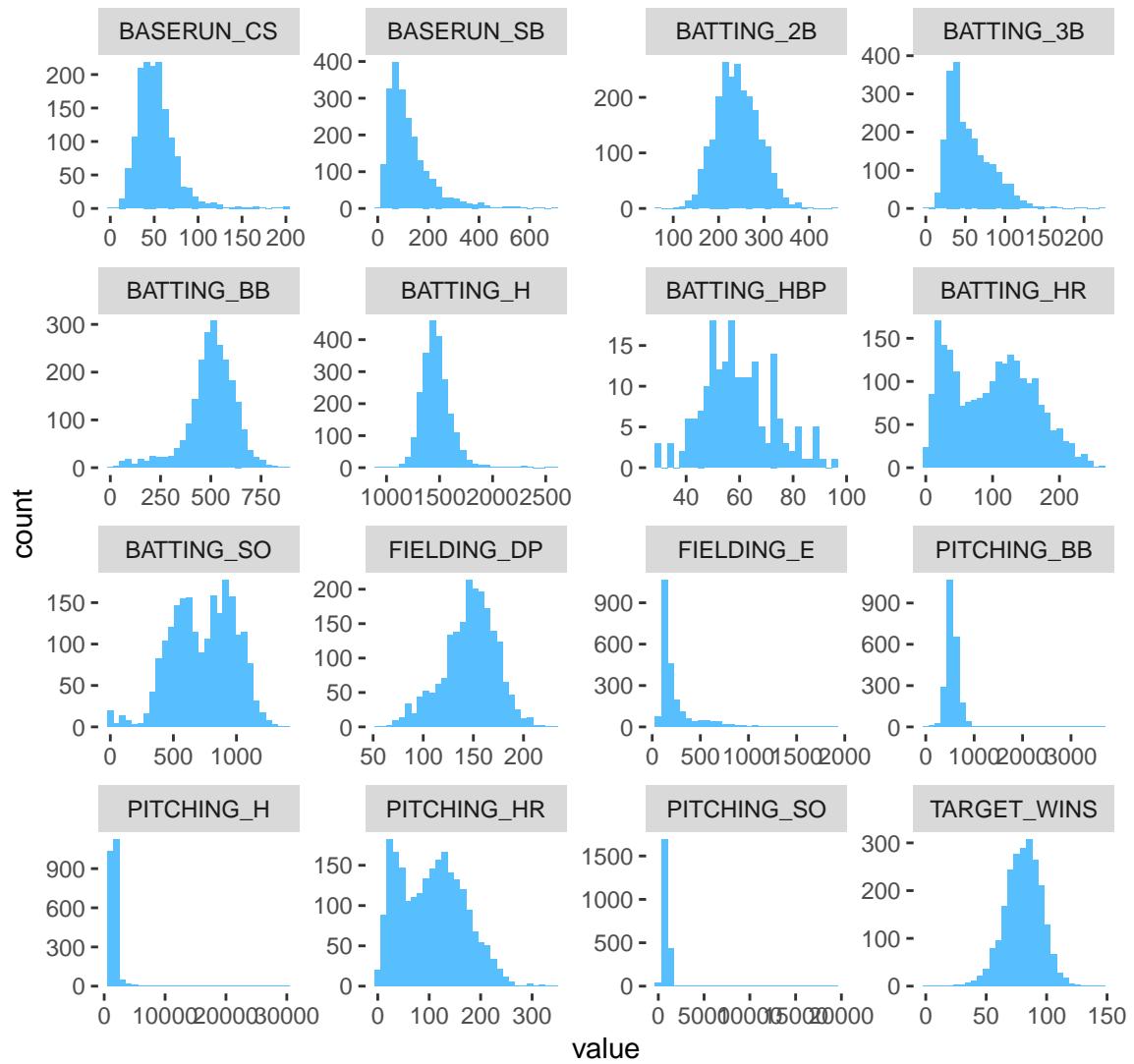
	n	min	mean	median	max	sd
TARGET_WINS	2276	0	80.79086	82.0	146	15.75215
BATTING_H	2276	891	1469.26977	1454.0	2554	144.59120
BATTING_2B	2276	69	241.24692	238.0	458	46.80141
BATTING_3B	2276	0	55.25000	47.0	223	27.93856
BATTING_HR	2276	0	99.61204	102.0	264	60.54687
BATTING_BB	2276	0	501.55888	512.0	878	122.67086
BATTING_SO	2174	0	735.60534	750.0	1399	248.52642
BASERUN_SB	2145	0	124.76177	101.0	697	87.79117
BASERUN_CS	1504	0	52.80386	49.0	201	22.95634
BATTING_HBP	191	29	59.35602	58.0	95	12.96712
PITCHING_H	2276	1137	1779.21046	1518.0	30132	1406.84293
PITCHING_HR	2276	0	105.69859	107.0	343	61.29875
PITCHING_BB	2276	0	553.00791	536.5	3645	166.35736
PITCHING_SO	2174	0	817.73045	813.5	19278	553.08503
FIELDING_E	2276	65	246.48067	159.0	1898	227.77097
FIELDING_DP	1990	52	146.38794	149.0	228	26.22639

## 1.5 Data Visualization

### 1.5.1 Shape of Predictor Distributions

The distribution of most of the variables seems normal although BASERUN\_SB, BASERUN\_CS, and BATTING\_3B have a slight to moderate right skew, FIELDING\_E, PITCHING\_BB, PITCHING\_H, and PITCHING\_SO have an extreme right skew, and BATTING\_HR, BATTING\_SO, and PITCHING\_HR are bimodal. As a result some data transformation will most likely improve the accuracy of our model.

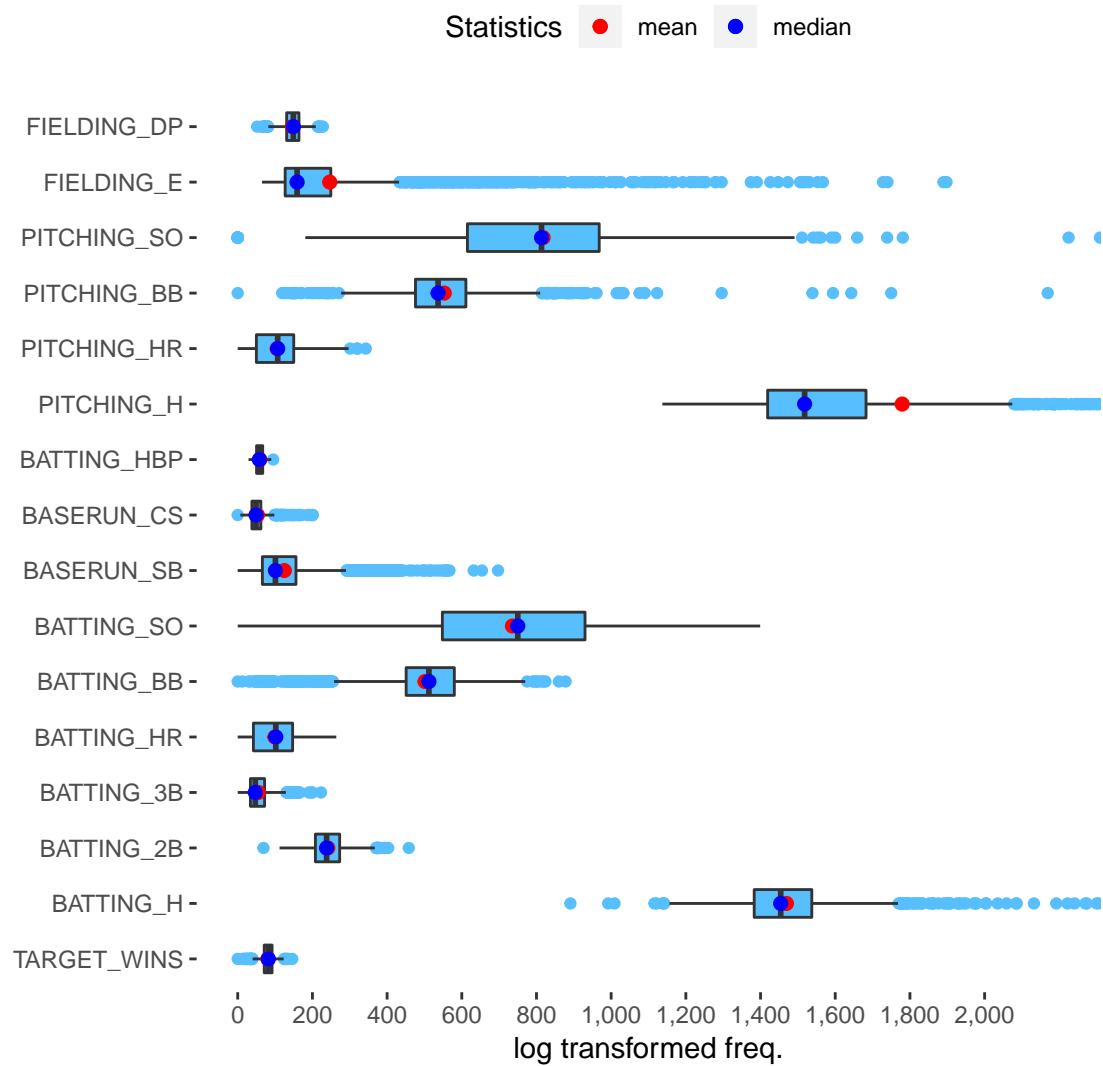
## Data Distributions



### 1.5.2 Outliers

There are also a large number of outliers that may need to be handled... (need to write more here...)

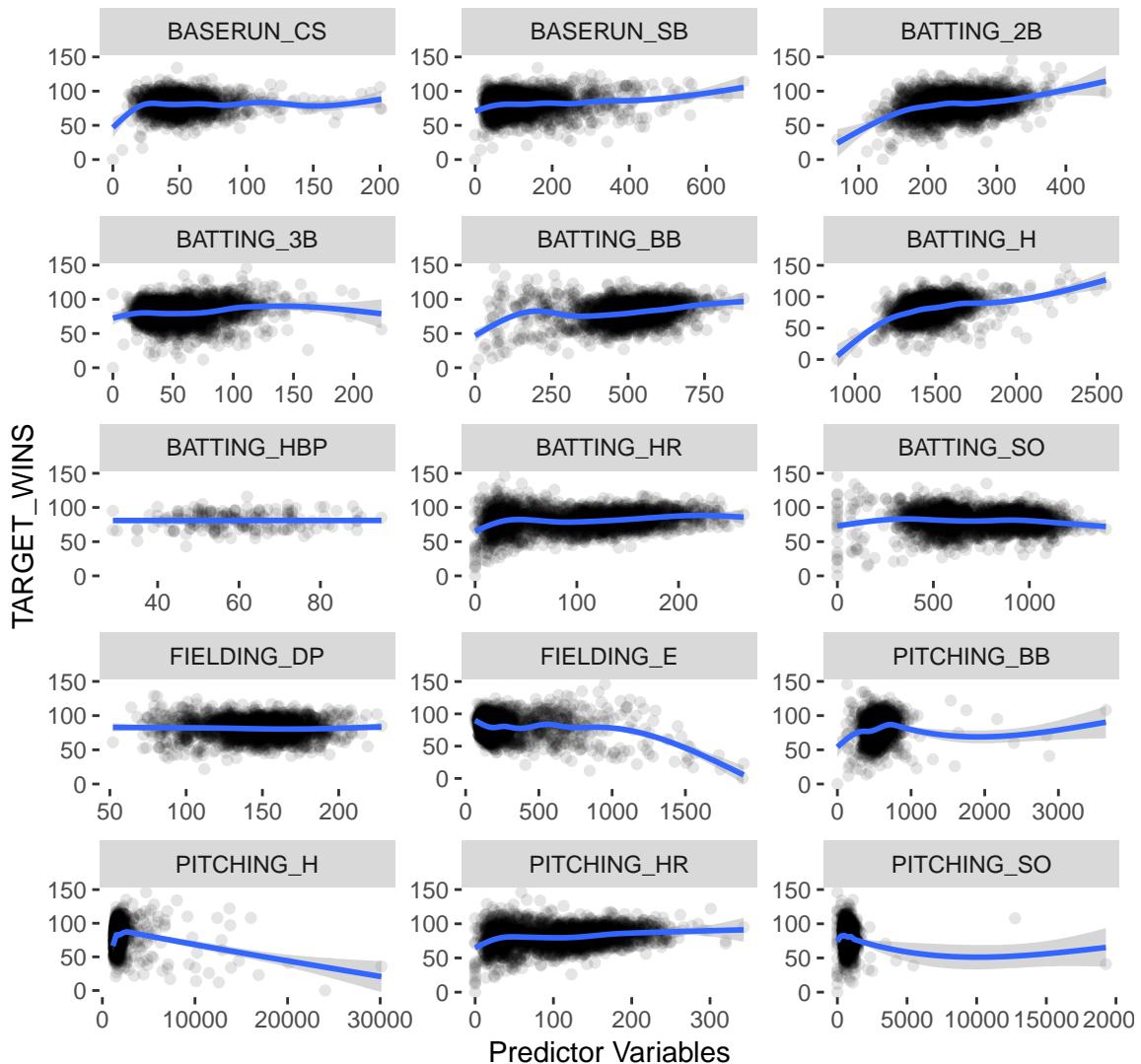
## Boxplot



### 1.5.3 Linearity

Need to write something here...

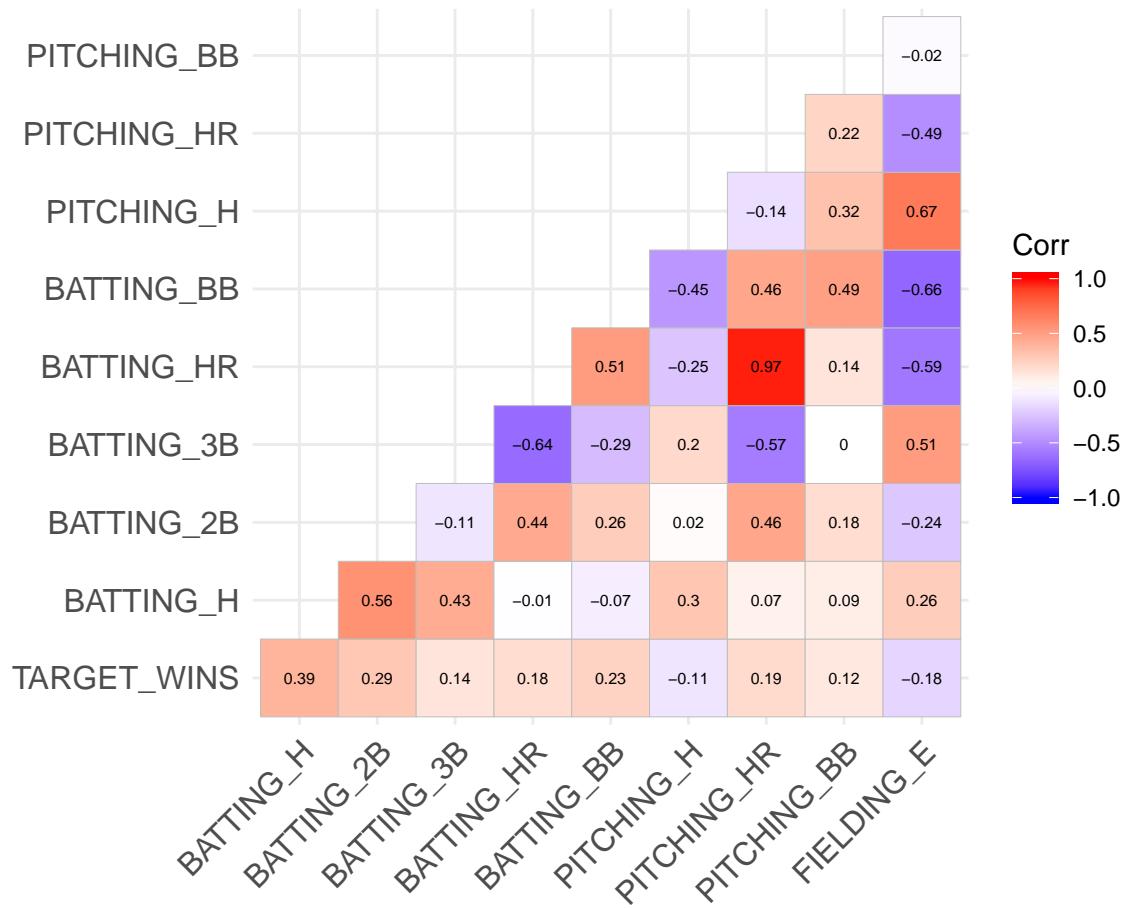
## Each Predictor vs. Target



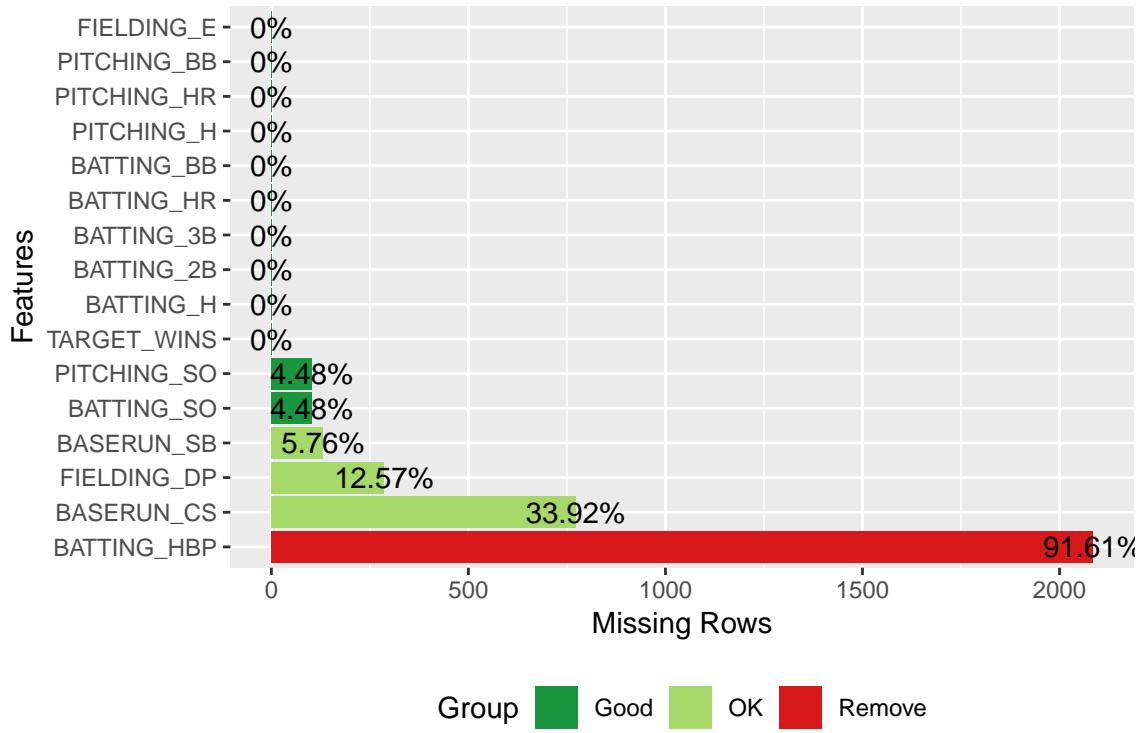
The theoretical effect of strikeouts by batters, batters caught stealing, errors, walks, hits, and homeruns allowed were believed to theoretically have a negative impact on the number of wins of an individual team in a given year. A closer look at the correlation between the variables painted a different picture.

When compared to what was hypothesized, there was actually a positive impact for the number of wins for a team in a given year by walks, hits, and homeruns allowed; at the same time, variables previously thought to have a positive correlation - strikeouts by pitchers and double plays - had a negative correlation for the number of wins. The three variables with the greatest correlation to the number of wins were the hits allowed, the walks by batters, and the walks allowed. Of these, the hits allowed had a relatively low correlation with the walks by batters and the walks allowed, whereas the walks allowed and the walks by batters had a direct positive correlation with one another.

Correlation



Of all the observations gathered across these fifteen variables, 10.187% were missing information; batters hit by pitches was missing the most, with 2,085 instances of missing information. The standard deviation of the various variables hints at the intense skewing of some of the variables provided, especially the hits allowed, walks allowed and strike outs.



## 2 DATA PREPARATION

### 2.1 Instructions:

Describe how you have transformed the data by changing the original variables or creating new variables. If you did transform the data or create new variables, discuss why you did this. Here are some possible transformations.

- Fix missing values (maybe with a Mean or Median value)
- Create flags to suggest if a variable was missing
- Transform data by putting it into buckets
- Mathematical transforms such as log or square root
- Combine variables (such as ratios or adding or multiplying) to create new variables

### 2.2 Possible writeup for Data Preparation

As previously mentioned, just north of 10% of the data was missing values. Missing values can lead to errors in a model, bias, and worse if left unaccounted for. Attempting to “fix” this by imputing values or guessing why the values are missing in the first place - such as concluding that the missing values are meant to be zeroes - are just as likely to help with creating a model as it is to help with creating a disaster.

One of the R packages utilized, DataExplorer, recommends removing null or missing values; it was for this reason all observations of hits by pitch were removed.<< SOURCE FOR THIS IS REQUIRED! IF NO SOURCE IS PROVIDED, CONSIDER USING “Due to the sheer volume of missing values present in the observations for hits by pitch (91.61%) it was determined the best course of action was to remove the variable altogether.” OR A VARIATION THEREOF. >> Deleting all cases with missing values, in this instance, would have shrunk the size of the dataset down to less than a tenth of its original size. For this reason, the feature itself was excluded from the dataset, rather than the cases that had no values present for it.

The other missing values - present in batting strikeouts... needs more work. x\_x

## 2.3 Information in general, don't use the writeups here as-is

## 2.4 Missing Values

- 1) Hit by pitch missing 91.61% .
  - Missing values can lead to errors and bias into a model. Fixing and imputation may help or make it worse.
  - When it is just a few observations missing, modifications can be made, however, with 91.61% is a large proportion and could distort the modelling later on that it is better to ignore this column.
  - The Data explorer package recommends to remove.
  - From LMR: Missing Completely at Random (MCAR) The probability that a value is missing is the same for all cases. If we simply delete all cases with missing values from the analysis, we will cause no bias, although we may lose some information.
  - However, there is no consensus on when to exclude missing data. Some argue that missing data more than 10% can lead to bias. Others argue that missing data patterns have greater impact than the proportion.
- 2) Pitching\_SO and Batting\_SO are missing exact same proportion 4.48% and are missing in the same observations.

## 2.5 NA Imputation

### 2.5.1 Data before imputing values

	n	min	mean	median	max	sd
TARGET_WINS	2276	0	80.79086	82.0	146	15.75215
BATTING_H	2276	891	1469.26977	1454.0	2554	144.59120
BATTING_2B	2276	69	241.24692	238.0	458	46.80141
BATTING_3B	2276	0	55.25000	47.0	223	27.93856
BATTING_HR	2276	0	99.61204	102.0	264	60.54687
BATTING_BB	2276	0	501.55888	512.0	878	122.67086
BATTING_SO	2174	0	735.60534	750.0	1399	248.52642
BASERUN_SB	2145	0	124.76177	101.0	697	87.79117
BASERUN_CS	1504	0	52.80386	49.0	201	22.95634
PITCHING_H	2276	1137	1779.21046	1518.0	30132	1406.84293
PITCHING_HR	2276	0	105.69859	107.0	343	61.29875
PITCHING_BB	2276	0	553.00791	536.5	3645	166.35736
PITCHING_SO	2174	0	817.73045	813.5	19278	553.08503
FIELDING_E	2276	65	246.48067	159.0	1898	227.77097
FIELDING_DP	1990	52	146.38794	149.0	228	26.22639

### 2.5.2 Data after imputing values

	n	min	mean	median	max	sd
TARGET_WINS	2276	0	80.79086	82.00000	146	15.75215
BATTING_H	2276	891	1469.26977	1454.00000	2554	144.59120
BATTING_2B	2276	69	241.24692	238.00000	458	46.80141
BATTING_3B	2276	0	55.25000	47.00000	223	27.93856
BATTING_HR	2276	0	99.61204	102.00000	264	60.54687
BATTING_BB	2276	0	501.55888	512.00000	878	122.67086
BATTING_SO	2276	0	728.22986	739.00000	1399	246.65040
BASERUN_SB	2276	0	124.62700	105.00000	697	85.28361
BASERUN_CS	2276	0	70.73189	56.42774	201	38.01002
PITCHING_H	2276	1137	1779.21046	1518.00000	30132	1406.84293
PITCHING_HR	2276	0	105.69859	107.00000	343	61.29875
PITCHING_BB	2276	0	553.00791	536.50000	3645	166.35736
PITCHING_SO	2276	0	807.62995	803.50000	19278	543.08096
FIELDING_E	2276	65	246.48067	159.00000	1898	227.77097
FIELDING_DP	2276	52	145.30987	146.00000	228	24.89909

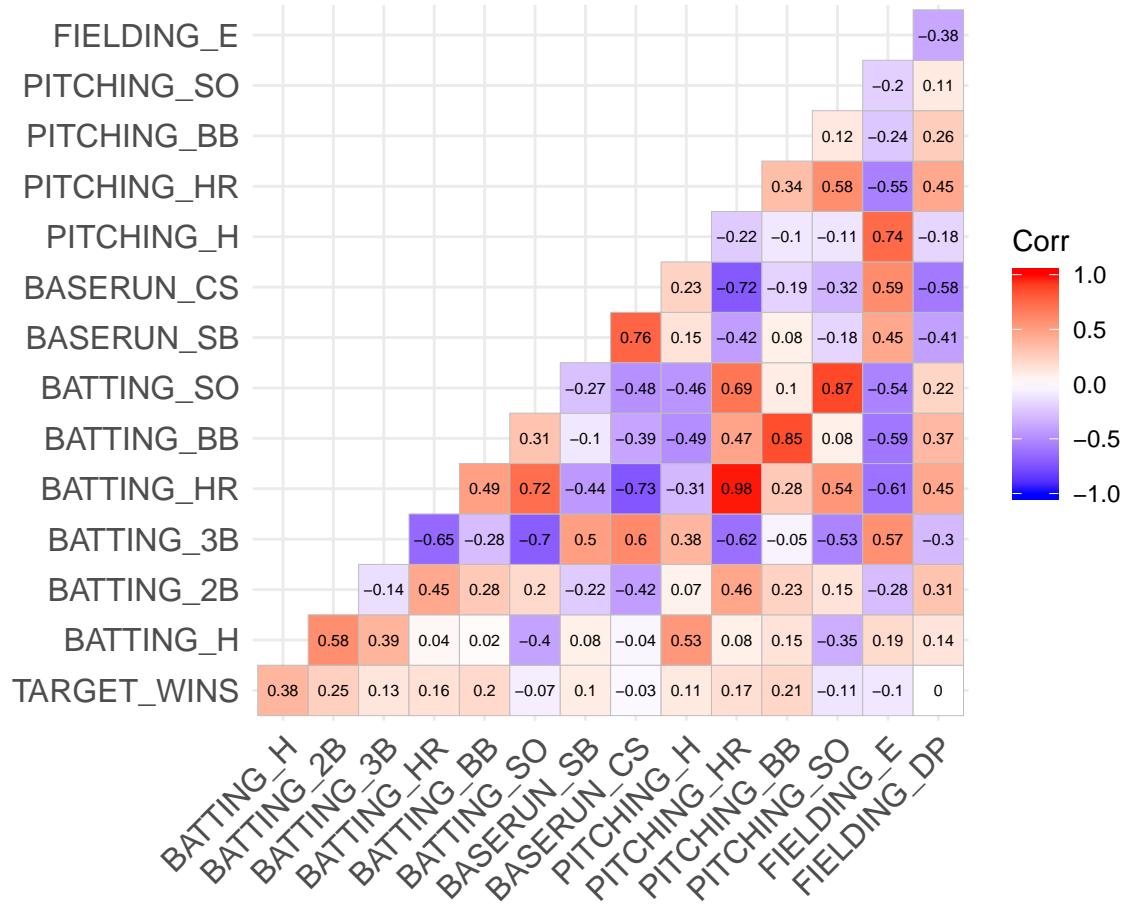
### 2.5.3 Difference between original and imputed data

	n	min	mean	median	max	sd
TARGET_WINS	0	0	0.0000000	0.000000	0	0.000000
BATTING_H	0	0	0.0000000	0.000000	0	0.000000
BATTING_2B	0	0	0.0000000	0.000000	0	0.000000
BATTING_3B	0	0	0.0000000	0.000000	0	0.000000
BATTING_HR	0	0	0.0000000	0.000000	0	0.000000
BATTING_BB	0	0	0.0000000	0.000000	0	0.000000
BATTING_SO	-102	0	7.3754714	11.000000	0	1.876015
BASERUN_SB	-131	0	0.1347715	-4.000000	0	2.507556
BASERUN_CS	-772	0	-17.9280286	-7.427742	0	-15.053684
PITCHING_H	0	0	0.0000000	0.000000	0	0.000000
PITCHING_HR	0	0	0.0000000	0.000000	0	0.000000
PITCHING_BB	0	0	0.0000000	0.000000	0	0.000000
PITCHING_SO	-102	0	10.1004968	10.000000	0	10.004072
FIELDING_E	0	0	0.0000000	0.000000	0	0.000000
FIELDING_DP	-286	0	1.0780691	3.000000	0	1.327290

## 2.6 Remove Outliers

Need to do writeup for this?

## Correlation



```

## TARGET_WINS    BATTING_H    BATTING_2B    BATTING_3B    BATTING_HR    BATTING_BB
##      0          0          0          0          0          0
##  BATTING_SO    BASERUN_SB   BASERUN_CS   PITCHING_H   PITCHING_HR   PITCHING_BB
##     102        131         772         0          0          0
##  PITCHING_SO   FIELDING_E  FIELDING_DP
##     102         0          286

```

## 2.7 Feature Engineering

Jeremy: Adjusted this to reflect offense (batting) minus defense (pitching). These arithmetically transformed offense / defense variables are linearly related with BATTING and PITCHING variables, so we can include one or the other in a model, but not both. Replacing original variables with these transforms did not improve  $R^2$  in a base case.

FOR THE OTHER HALF OF THE GROUP:

```

z_train <- sapply(imputed_train, scale)
log_train <- log(imputed_train) # weird results
z_log_train <- sapply(log_train, scale) # weirder results

```

imputed\_train is most likely the variable you want to use.

### 3 BUILD MODELS

#### 3.1 Instructions:

Using the training data set, build at least three different multiple linear regression models, using different variables (or the same variables with different transformations). Since we have not yet covered automated variable selection methods, you should select the variables manually (unless you previously learned Forward or Stepwise selection, etc.). Since you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done. Discuss the coefficients in the models, do they make sense? For example, if a team hits a lot of Home Runs, it would be reasonably expected that such a team would win more games. However, if the coefficient is negative (suggesting that the team would lose more games), then that needs to be discussed. Are you keeping the model even though it is counter intuitive? Why? The boss needs to know.

#### 3.2 MODEL 1

Multiple regression can be created as a purely statistical model, through the use of significance tests, or it can be interpreted in a more practical, non-statistical manner. This approach is based on the subject-area expertise.

We've created the following categories from the most important to the least important variables according to the subject-area expert.

Very Important: BATTING\_H, BATTING\_HR, BATTING\_SO ,FIELDING\_E, PITCHING\_SO

Fairly Important: BASERUN\_SB, PITCHING\_HR, BATTING\_BB

Important: BATTING\_2B, BATTING\_3B, FIELDING\_DP, PITCHING\_H

Slightly Important: PITCHING\_BB, BASERUN\_CS

Not at all important: BATTING\_HBP

'Batters hit by pitch' and 'Caught Stealing' have been eliminated as least important variables according to the expert.

```
##  
## Call:  
## lm(formula = TARGET_WINS ~ BATTING_H + BATTING_HR + BATTING_SO +  
##      FIELDING_E + PITCHING_SO + BASERUN_SB + PITCHING_HR + BATTING_BB +  
##      BATTING_2B + BATTING_3B + FIELDING_DP + PITCHING_BB + PITCHING_H,  
##      data = imputed_train)  
##  
## Residuals:  
##      Min        1Q    Median        3Q       Max  
## -81.602   -8.272    0.065    7.985   69.047  
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 36.592272  5.663910  6.461 1.28e-10 ***  
## BATTING_H    0.016719  0.005010  3.337 0.000861 ***  
## BATTING_HR   0.119188  0.050659  2.353 0.018723 *  
## BATTING_SO   -0.031132  0.006598 -4.718 2.53e-06 ***  
## FIELDING_E   -0.038259  0.003433 -11.144 < 2e-16 ***  
## PITCHING_SO  0.019966  0.005420  3.684 0.000235 ***  
## BASERUN_SB   0.041446  0.004934  8.399 < 2e-16 ***
```

```

## PITCHING_HR -0.034985  0.046876 -0.746 0.455549
## BATTING_BB   0.078335  0.016217  4.830 1.46e-06 ***
## BATTING_2B   -0.011063  0.009390 -1.178 0.238836
## BATTING_3B    0.126375  0.018155  6.961 4.43e-12 ***
## FIELDING_DP  -0.091226  0.013111 -6.958 4.53e-12 ***
## PITCHING_BB  -0.055706  0.014380 -3.874 0.000110 ***
## PITCHING_H    0.013513  0.001539  8.779 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.63 on 2216 degrees of freedom
## Multiple R-squared:  0.2835, Adjusted R-squared:  0.2793
## F-statistic: 67.46 on 13 and 2216 DF,  p-value: < 2.2e-16

```

We got 0.2793 on Adjusted R-squared after we removed these two variables. Once we tried to remove other not very important variables according to subject-area expert, we got an even lower R-squared.

The next step we performed was backward elimination, which was more effective compared to forward selection. BATTING\_H and BATTING\_2B have been removed based on the Backward Selection results.

```

## Start:  AIC=11325.02
## TARGET_WINS ~ BATTING_H + BATTING_HR + BATTING_SO + FIELDING_E +
##               PITCHING_SO + BASERUN_SB + PITCHING_HR + BATTING_BB + BATTING_2B +
##               BATTING_3B + FIELDING_DP + PITCHING_BB + PITCHING_H
##
##          Df Sum of Sq   RSS   AIC
## - PITCHING_HR  1     88.9 353605 11324
## - BATTING_2B   1    221.5 353737 11324
## <none>           353516 11325
## - BATTING_HR   1    883.1 354399 11329
## - BATTING_H    1   1776.5 355292 11334
## - PITCHING_SO  1   2165.1 355681 11337
## - PITCHING_BB  1   2393.9 355910 11338
## - BATTING_SO   1   3551.6 357068 11345
## - BATTING_BB   1   3722.3 357238 11346
## - FIELDING_DP  1   7723.5 361239 11371
## - BATTING_3B   1   7730.1 361246 11371
## - BASERUN_SB   1  11255.0 364771 11393
## - PITCHING_H   1  12294.0 365810 11399
## - FIELDING_E   1  19811.3 373327 11445
##
## Step:  AIC=11323.58
## TARGET_WINS ~ BATTING_H + BATTING_HR + BATTING_SO + FIELDING_E +
##               PITCHING_SO + BASERUN_SB + BATTING_BB + BATTING_2B + BATTING_3B +
##               FIELDING_DP + PITCHING_BB + PITCHING_H
##
##          Df Sum of Sq   RSS   AIC
## - BATTING_2B   1    218.7 353824 11323
## <none>           353605 11324
## - BATTING_H    1   1790.1 355395 11333
## - PITCHING_SO  1   2131.4 355736 11335
## - BATTING_SO   1   3491.5 357096 11344
## - PITCHING_BB  1   4558.7 358164 11350
## - BATTING_BB   1   6449.9 360055 11362
## - BATTING_3B   1   7663.0 361268 11369

```

```

## - FIELDING_DP 1 7822.8 361428 11370
## - BATTING_HR 1 11191.6 364796 11391
## - BASERUN_SB 1 11302.9 364908 11392
## - PITCHING_H 1 12384.3 365989 11398
## - FIELDING_E 1 20640.3 374245 11448
##
## Step: AIC=11322.96
## TARGET_WINS ~ BATTING_H + BATTING_HR + BATTING_SO + FIELDING_E +
##      PITCHING_SO + BASERUN_SB + BATTING_BB + BATTING_3B + FIELDING_DP +
##      PITCHING_BB + PITCHING_H
##
##          Df Sum of Sq    RSS   AIC
## <none>             353824 11323
## - BATTING_H 1 1777.1 355601 11332
## - PITCHING_SO 1 1990.8 355814 11334
## - BATTING_SO 1 3479.4 357303 11343
## - PITCHING_BB 1 4495.2 358319 11349
## - BATTING_BB 1 6351.2 360175 11361
## - FIELDING_DP 1 7773.8 361597 11369
## - BATTING_3B 1 8094.2 361918 11371
## - BATTING_HR 1 11452.8 365276 11392
## - BASERUN_SB 1 11828.7 365652 11394
## - PITCHING_H 1 12901.1 366725 11401
## - FIELDING_E 1 20437.9 374261 11446
##
## Call:
## lm(formula = TARGET_WINS ~ BATTING_H + BATTING_HR + BATTING_SO +
##      FIELDING_E + PITCHING_SO + BASERUN_SB + BATTING_BB + BATTING_3B +
##      FIELDING_DP + PITCHING_BB + PITCHING_H, data = imputed_train)
##
## Coefficients:
## (Intercept) BATTING_H BATTING_HR BATTING_SO FIELDING_E
## 38.93008 0.01316 0.08286 -0.03074 -0.03727
## PITCHING_SO BASERUN_SB BATTING_BB BATTING_3B FIELDING_DP
## 0.01896 0.04219 0.08441 0.12819 -0.09140
## PITCHING_BB PITCHING_H
## -0.06157 0.01375
##
## Call:
## lm(formula = TARGET_WINS ~ BATTING_HR + BATTING_SO + FIELDING_E +
##      PITCHING_SO + BASERUN_SB + PITCHING_HR + BATTING_BB + BATTING_3B +
##      FIELDING_DP + PITCHING_BB + PITCHING_H, data = imputed_train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -92.349 -8.319  0.095  8.044  73.821
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 50.967361  3.815685 13.357 < 2e-16 ***
## BATTING_HR  0.137603  0.050478  2.726 0.006461 **
## BATTING_SO -0.033187  0.006582 -5.042 4.98e-07 ***
## FIELDING_E -0.041649  0.003245 -12.835 < 2e-16 ***

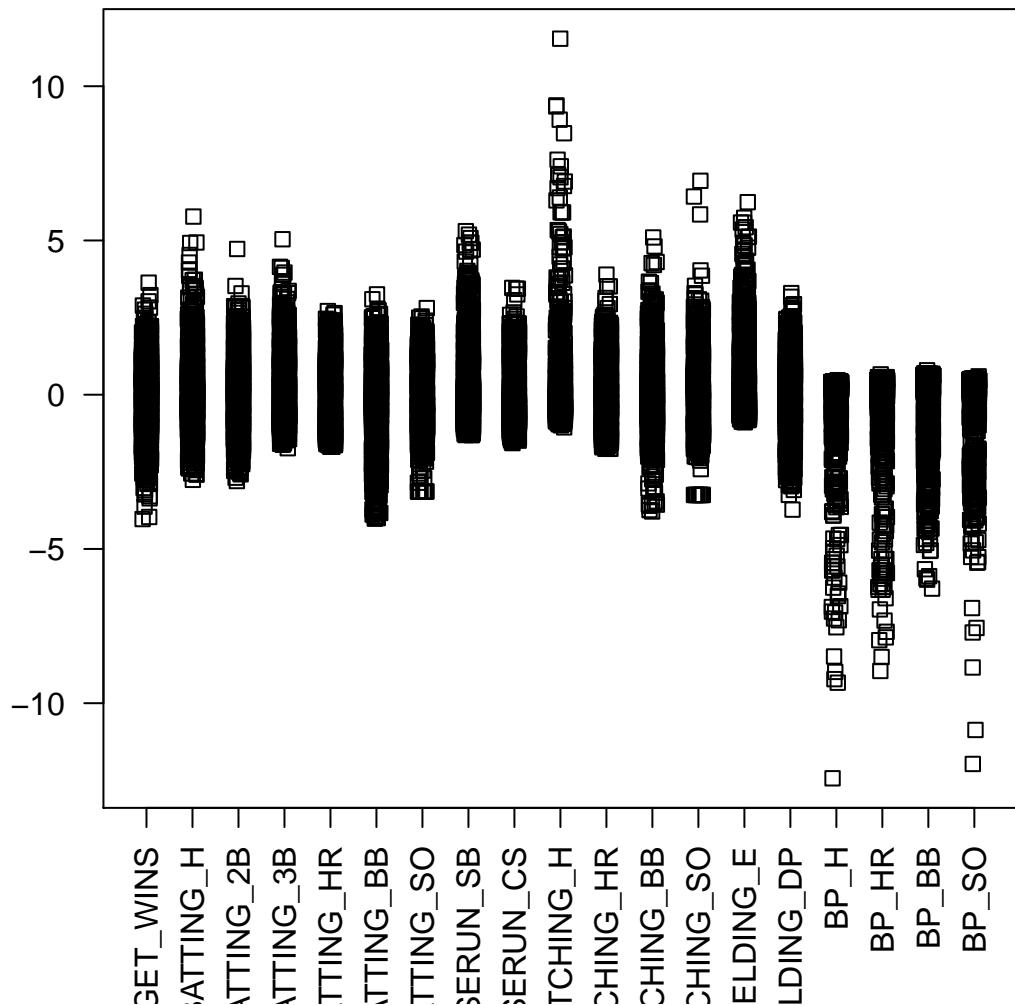
```

```

## PITCHING_SO  0.018692  0.005384   3.472 0.000527 ***
## BASERUN_SB   0.045791  0.004785   9.569 < 2e-16 ***
## PITCHING_HR -0.038873  0.046964  -0.828 0.407915
## BATTING_BB   0.085893  0.016055   5.350 9.70e-08 ***
## BATTING_3B   0.156437  0.016039   9.754 < 2e-16 ***
## FIELDING_DP  -0.083573  0.012950  -6.454 1.34e-10 ***
## PITCHING_BB  -0.061960  0.014270  -4.342 1.48e-05 ***
## PITCHING_H    0.016931  0.001185  14.286 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.66 on 2218 degrees of freedom
## Multiple R-squared:  0.2795, Adjusted R-squared:  0.276
## F-statistic: 78.23 on 11 and 2218 DF,  p-value: < 2.2e-16

```

Our R-squared was still low (0.276), so we decided to look at the outliers, which can affect our model. Pitching\_h had the high number of outliers which indicated a need for data transformation. We decided to use log transformation for this variable.



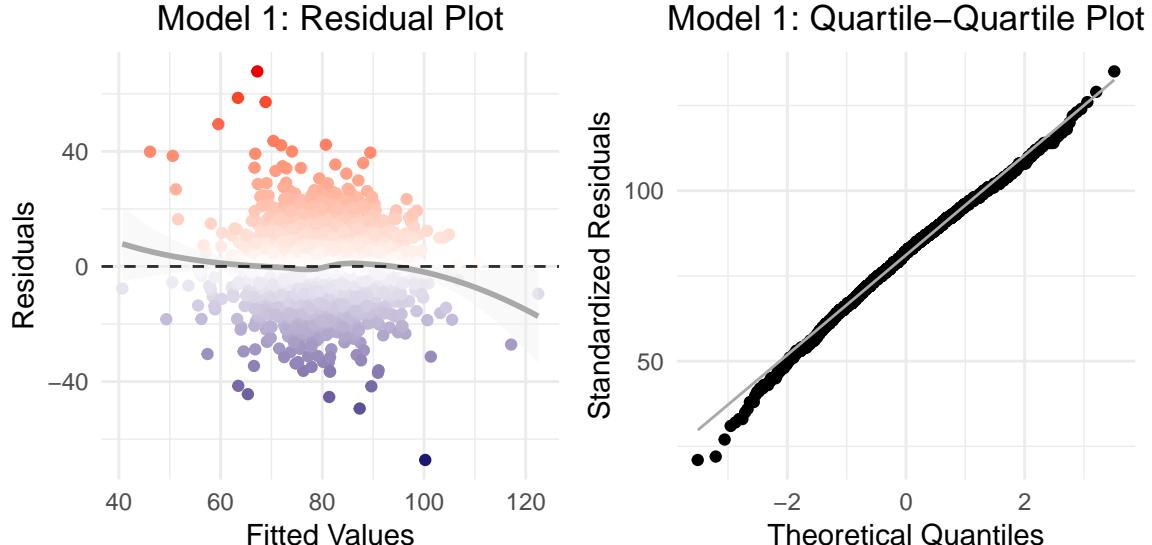
```
##  
## Call:  
## lm(formula = TARGET_WINS ~ BATTING_HR + BATTING_SO + FIELDING_E +  
##      PITCHING_SO + BASERUN_SB + PITCHING_HR + BATTING_BB + BATTING_3B +
```

```

##      FIELDING_DP + PITCHING_BB + log(PITCHING_H) , data = imputed_train)
##
## Residuals:
##      Min       1Q   Median     3Q    Max
## -67.238 -8.139   0.185  8.020 67.777
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -3.037e+02  2.426e+01 -12.522 < 2e-16 ***
## BATTING_HR            1.134e-01  5.003e-02   2.268  0.0234 *
## BATTING_SO            -1.544e-02  6.668e-03  -2.315  0.0207 *
## FIELDING_E            -4.202e-02  3.159e-03 -13.299 < 2e-16 ***
## PITCHING_SO            7.688e-03  5.336e-03   1.441  0.1498
## BASERUN_SB             4.138e-02  4.680e-03   8.842 < 2e-16 ***
## PITCHING_HR            -4.352e-02  4.640e-02  -0.938  0.3484
## BATTING_BB             1.011e-01  1.600e-02   6.319 3.16e-10 ***
## BATTING_3B             1.171e-01  1.594e-02   7.349 2.79e-13 ***
## FIELDING_DP            -9.430e-02  1.283e-02  -7.351 2.76e-13 ***
## PITCHING_BB            -7.559e-02  1.423e-02  -5.311 1.20e-07 ***
## log(PITCHING_H)        5.222e+01  3.243e+00  16.099 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.52 on 2218 degrees of freedom
## Multiple R-squared:  0.2956, Adjusted R-squared:  0.2921
## F-statistic:  84.6 on 11 and 2218 DF, p-value: < 2.2e-16

```

After we used the log trasformation our model's Adjusted R-squared increased to 0.2921.



```

##      Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
## 60.82 76.09 81.12 81.44 86.10 108.26 54

```

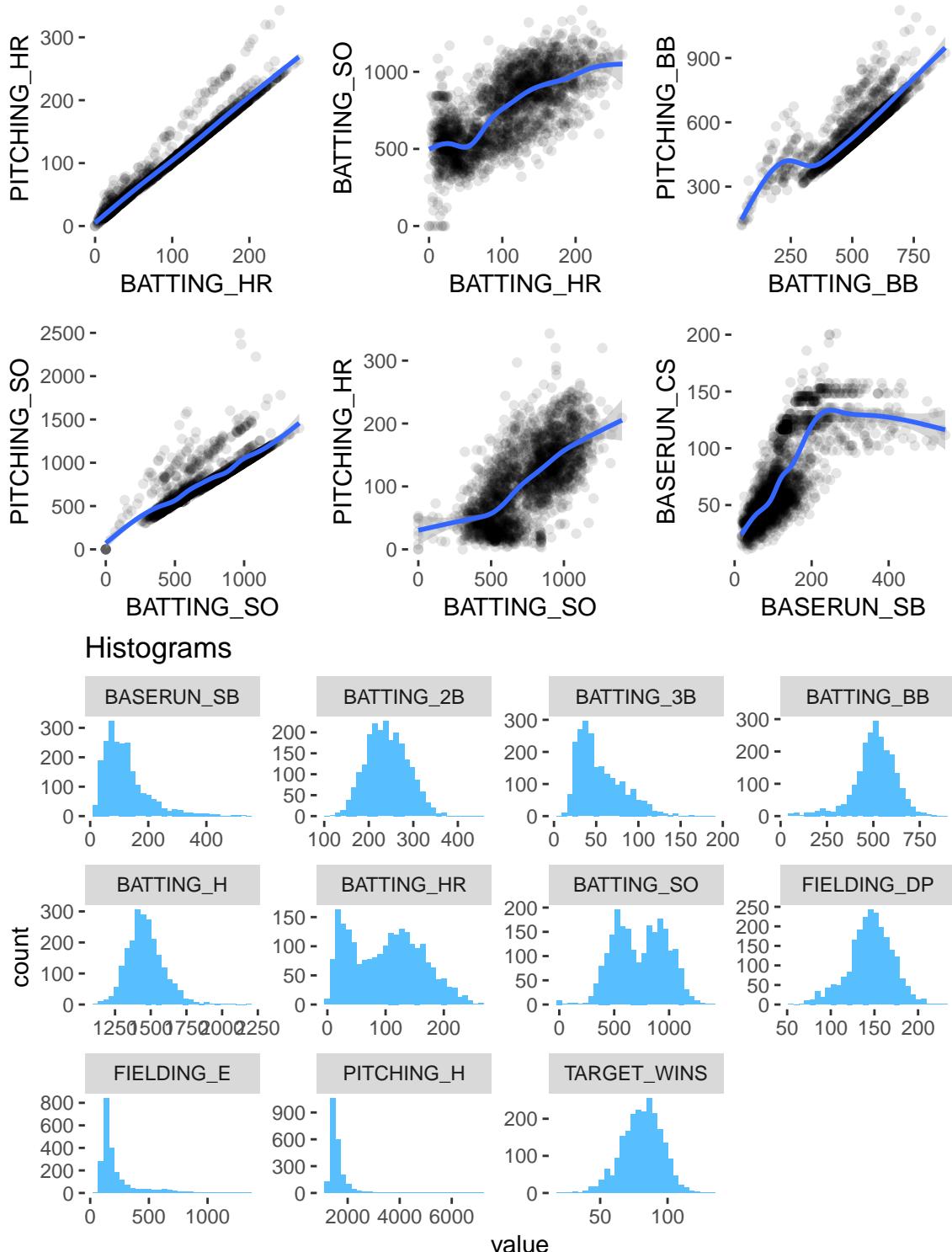
	r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.
value	0.2955663	0.2920727	12.51838	84.6026	0	12	-8793.869	17613.74	17687.96	347582.5	

Summary of Results for Model 1:

The overall Subject-Area expertise wasn't as effective as a stand alone method of creating multiple regression

models. Statistical iterations which were performed contradicted the subject area expert, such as, removing Batting\_H from the model. Additionally log transformation of PITCHING\_H made a significant improvement in our model linearity.

### 3.3 MODEL 2



### Each Predictor vs. Target

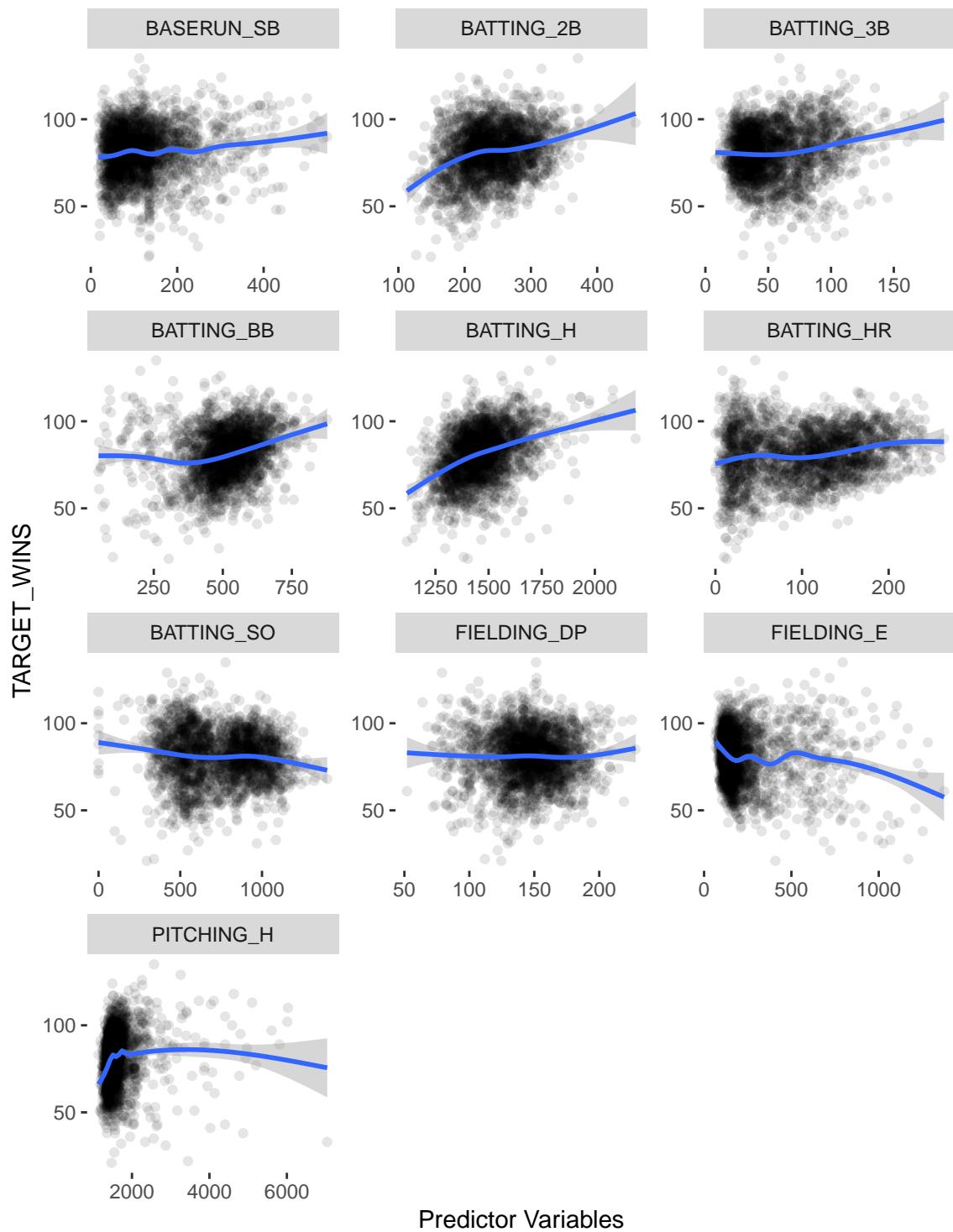
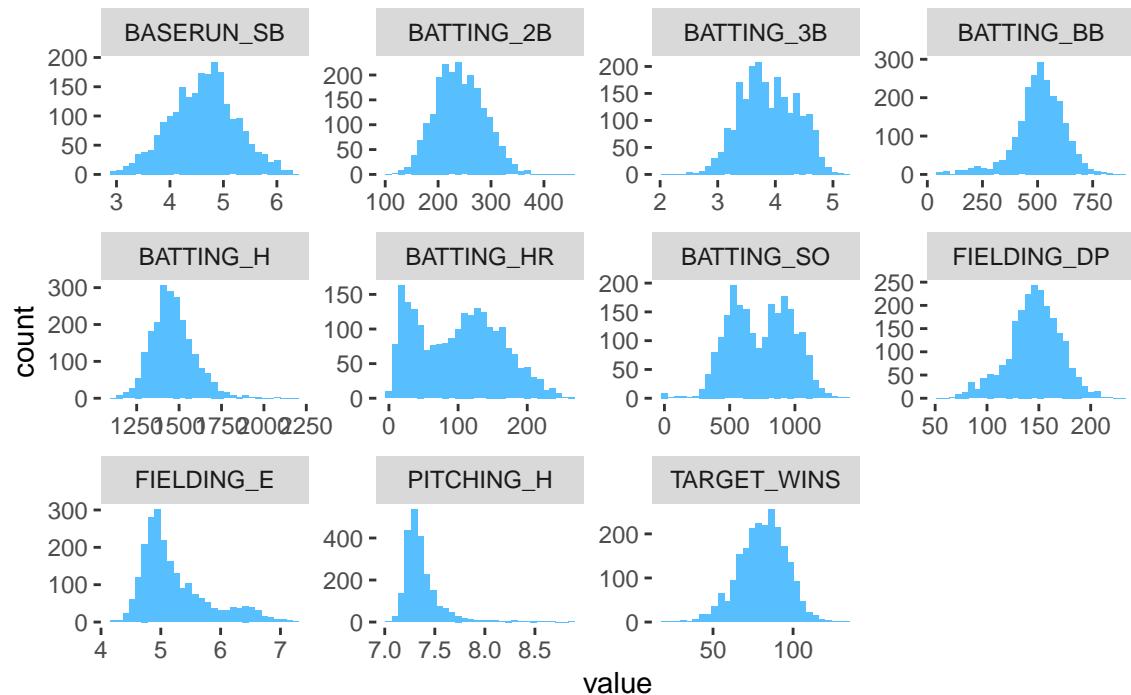


Figure 1: Each Predictor vs. Target

### 3.3.1 Log Transformed data

Histograms



```
##
## Call:
## lm(formula = TARGET_WINS ~ ., data = lm_data)
##
## Residuals:
##    Min     1Q   Median     3Q    Max 
## -52.645 -8.088   0.030   8.174  67.177 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -62.972452  15.913904 -3.957 7.83e-05 ***
## BATTING_H     0.025562   0.004611   5.543 3.32e-08 ***
## BATTING_2B   -0.029760   0.009216  -3.229  0.00126 ** 
## BATTING_3B     6.969781   0.947655   7.355 2.68e-13 ***
## BATTING_HR    0.070328   0.010126   6.946 4.93e-12 ***
## BATTING_BB    0.019229   0.003187   6.033 1.88e-09 ***
## BATTING_SO   -0.011228   0.002411  -4.656 3.41e-06 ***
## BASERUN_SB    4.546008   0.561766   8.092 9.54e-16 ***
## PITCHING_H   19.140858   2.616154   7.316 3.54e-13 ***
## FIELDING_E   -13.102762   1.061861 -12.339 < 2e-16 ***
## FIELDING_DP   -0.106723   0.013138  -8.123 7.47e-16 *** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 12.57 on 2219 degrees of freedom
## Multiple R-squared:  0.2889, Adjusted R-squared:  0.2857 
## F-statistic: 90.14 on 10 and 2219 DF,  p-value: < 2.2e-16
```

### Each Predictor vs. Target

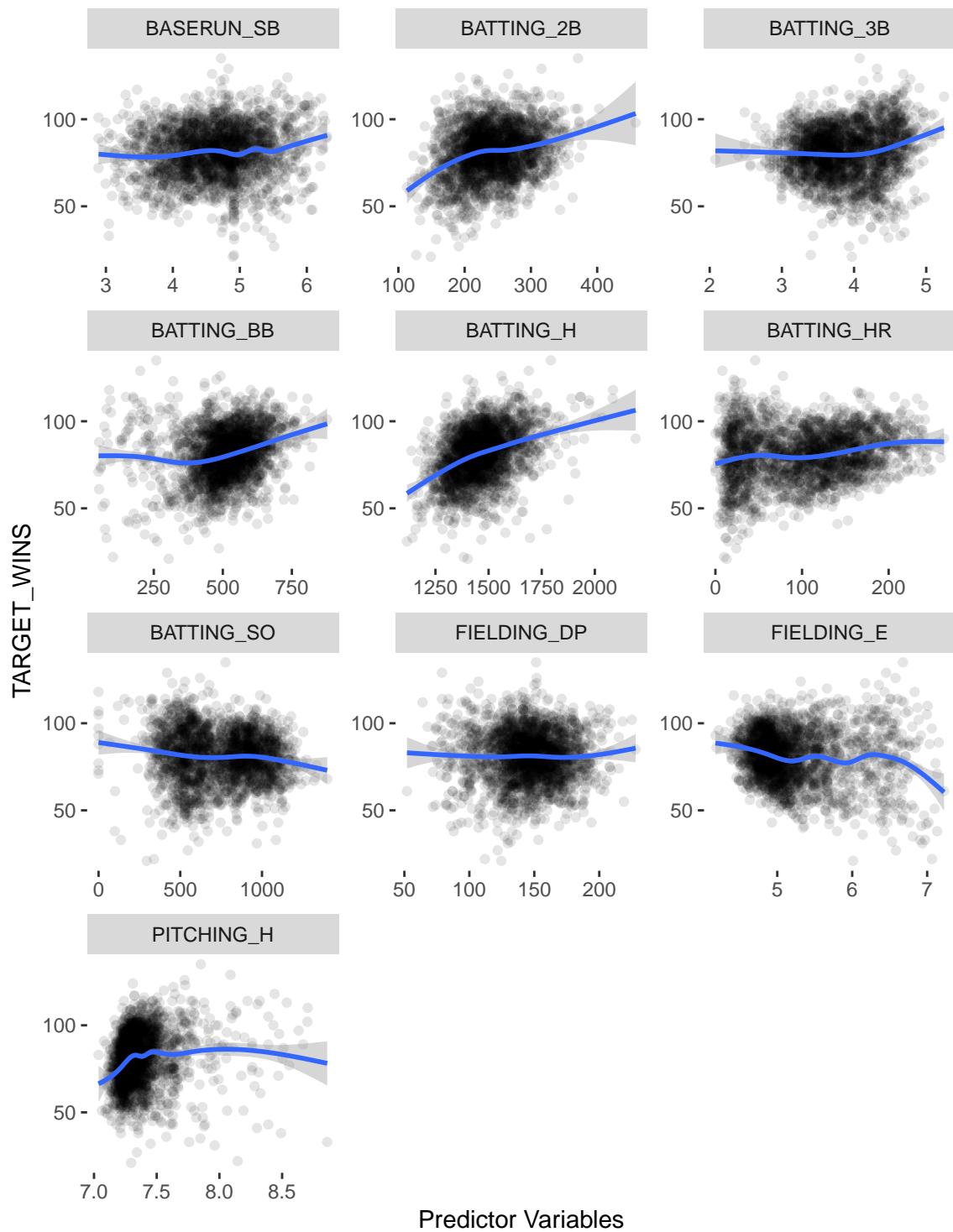
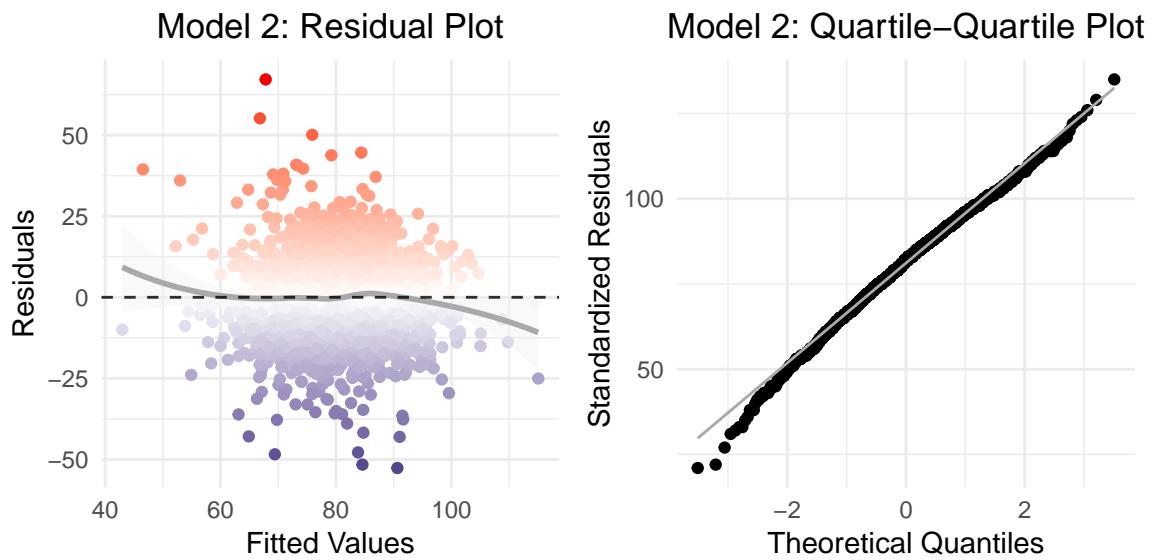
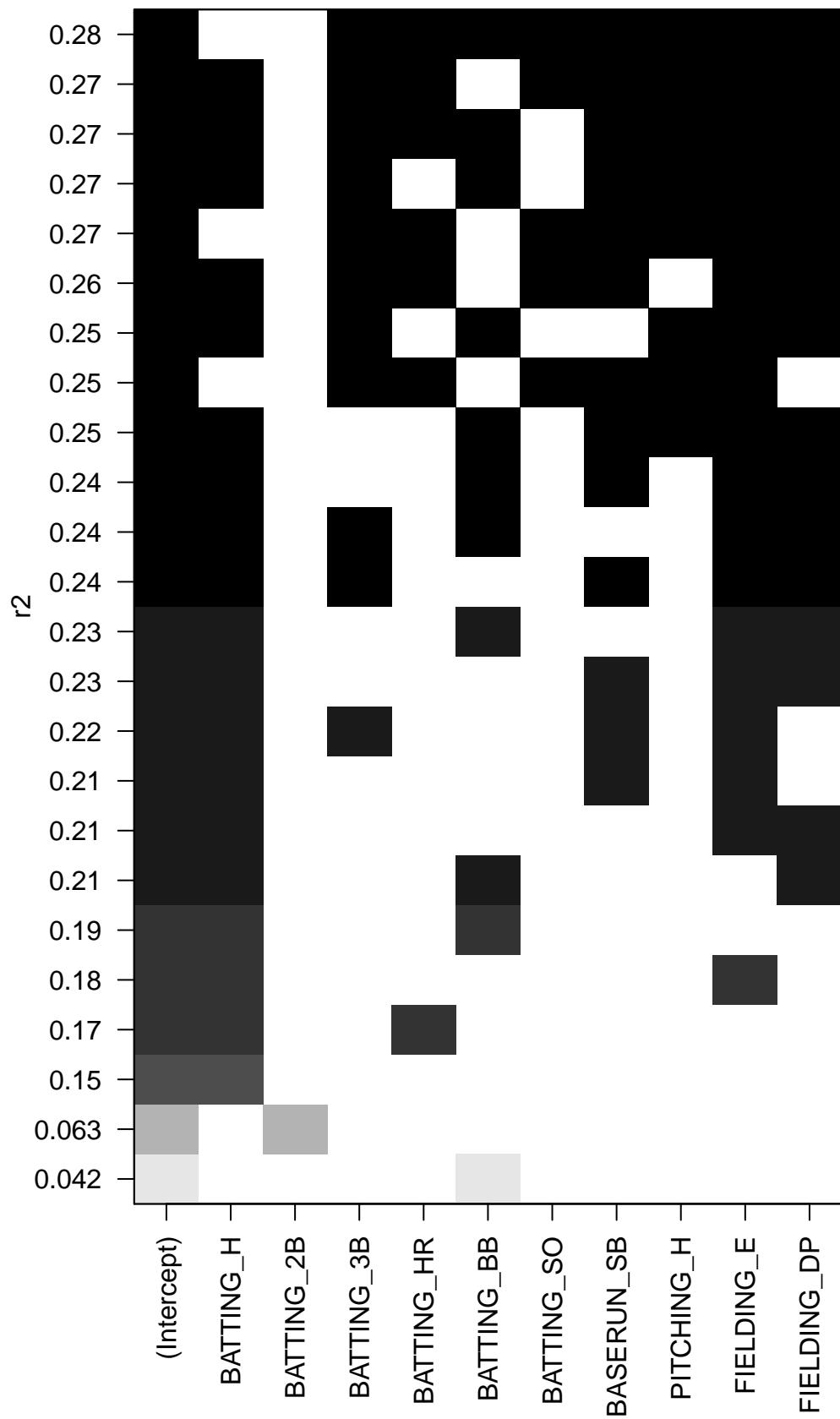


Figure 2: Log Transformed Predictors vs. Target

Table 1: Full Model Coefficients

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-62.9724524	15.9139042	-3.957071	0.0000783
BATTING_H	0.0255615	0.0046111	5.543426	0.0000000
BATTING_2B	-0.0297598	0.0092155	-3.229309	0.0012590
BATTING_3B	6.9697812	0.9476551	7.354765	0.0000000
BATTING_HR	0.0703282	0.0101256	6.945579	0.0000000
BATTING_BB	0.0192290	0.0031872	6.033192	0.0000000
BATTING_SO	-0.0112281	0.0024113	-4.656450	0.0000034
BASERUN_SB	4.5460080	0.5617657	8.092356	0.0000000
PITCHING_H	19.1408576	2.6161541	7.316411	0.0000000
FIELDING_E	-13.1027616	1.0618606	-12.339437	0.0000000
FIELDING_DP	-0.1067225	0.0131384	-8.122924	0.0000000



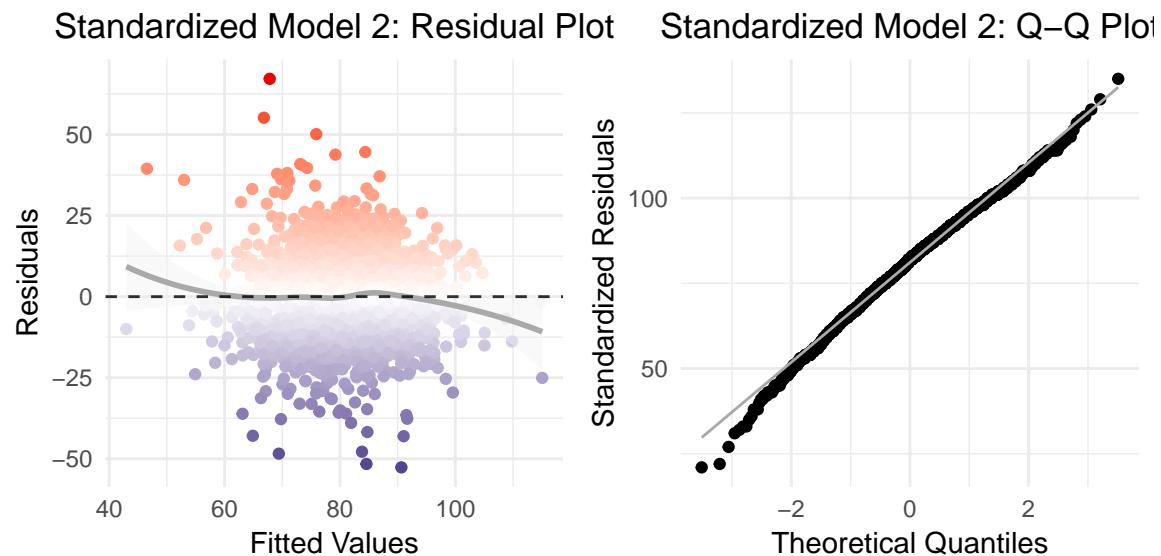


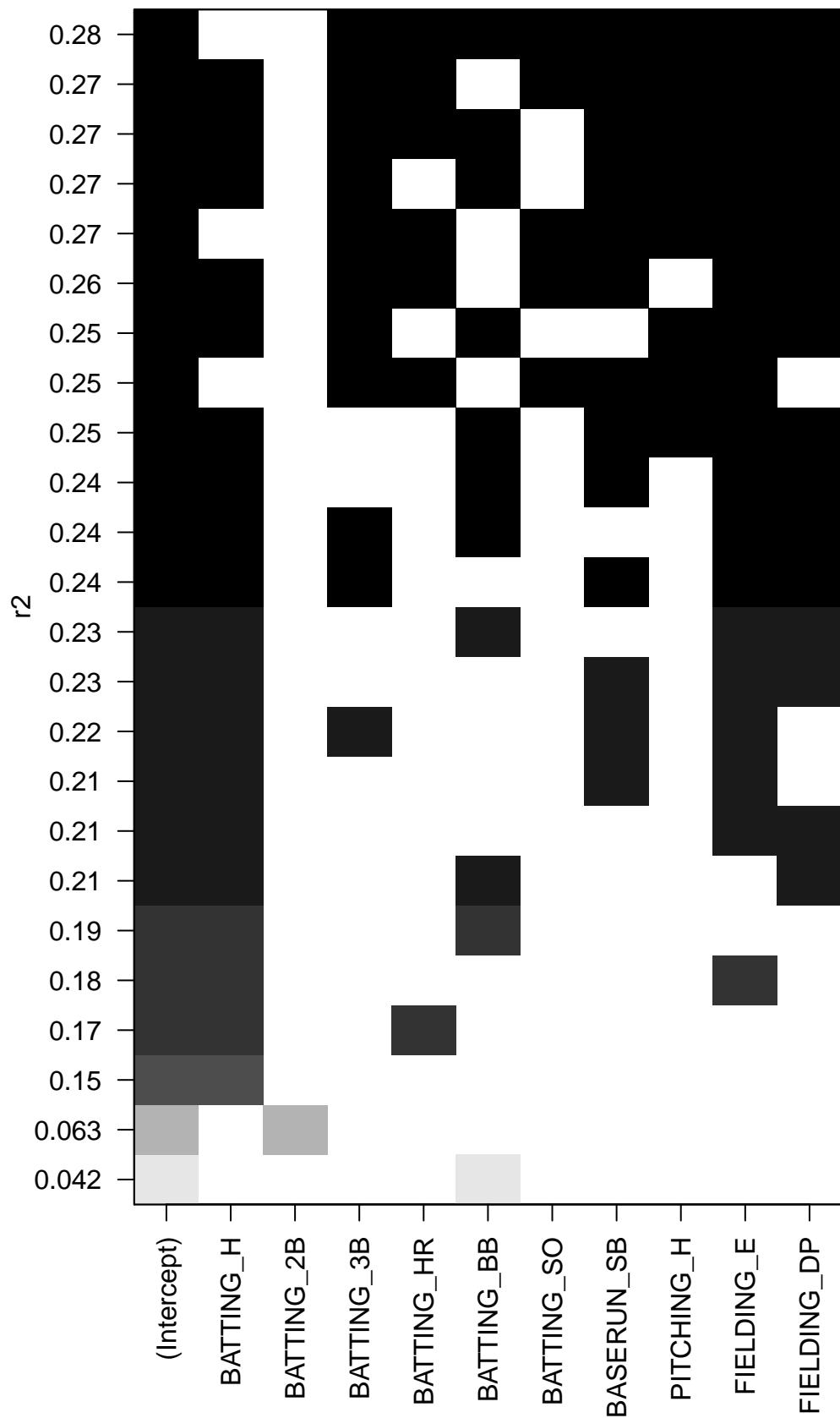
## NULL

Table 2: Full SCALED Model Coefficients

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	80.971749	0.2662860	304.078090	0.0000000
BATTING_H	3.222999	0.5814092	5.543426	0.0000000
BATTING_2B	-1.364857	0.4226469	-3.229309	0.0012590
BATTING_3B	3.381556	0.4597775	7.354765	0.0000000
BATTING_HR	4.220294	0.6076231	6.945579	0.0000000
BATTING_BB	2.190025	0.3629961	6.033192	0.0000000
BATTING_SO	-2.641712	0.5673232	-4.656450	0.0000034
BASERUN_SB	2.794642	0.3453435	8.092356	0.0000000
PITCHING_H	3.870518	0.5290186	7.316411	0.0000000
FIELDING_E	-7.441986	0.6031058	-12.339437	0.0000000
FIELDING_DP	-2.676186	0.3294610	-8.122924	0.0000000

```
## [1] 0.2888829
```





## NULL

### 3.3.2 Test all of the predictors

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
2229	493421.2	NA	NA	NA	NA
2219	350880.3	10	142541	90.14425	0

### 3.3.3 Test one predictor

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
2220	355739.4	NA	NA	NA	NA
2219	350880.3	1	4859.127	30.72958	0

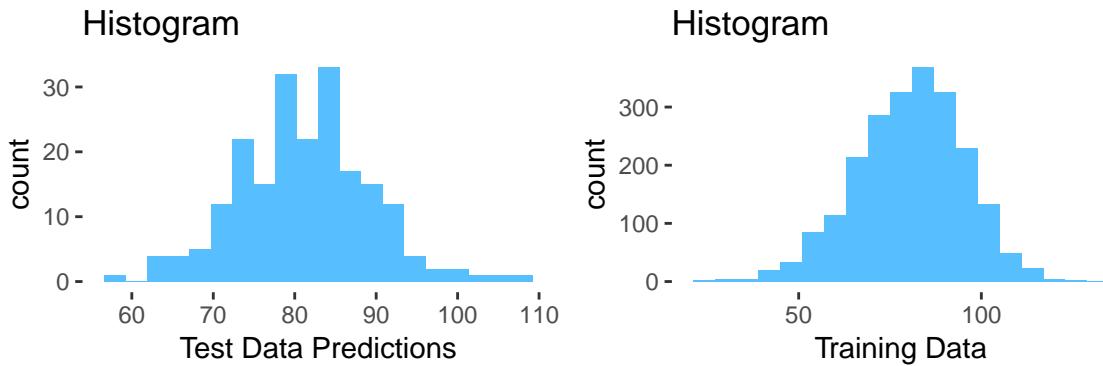
Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
2220	352529.3	NA	NA	NA	NA
2219	350880.3	1	1649.001	10.42844	0.001259

### 3.3.4 Testing a subspace

```

## 
## Call:
## lm(formula = TARGET_WINS ~ I(BATTING_HR + PITCHING_HR) + I(BATTING_BB +
##     PITCHING_BB) + I(BATTING_SO + PITCHING_SO) + BATTING_H +
##     BATTING_2B + BATTING_3B + BASERUN_SB + BASERUN_CS + PITCHING_H +
##     FIELDING_E + FIELDING_DP, data = all_data)
## 
## Residuals:
##      Min        1Q    Median        3Q       Max
## -67.298   -8.254    0.184    8.251   66.950
## 
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            33.769875  5.912262  5.712 1.27e-08 ***
## I(BATTING_HR + PITCHING_HR) 0.042078  0.004928  8.538 < 2e-16 ***
## I(BATTING_BB + PITCHING_BB) 0.006255  0.001547  4.044 5.44e-05 ***
## I(BATTING_SO + PITCHING_SO) -0.003880  0.001061 -3.656 0.000262 ***
## BATTING_H                0.023533  0.004888  4.815 1.57e-06 ***
## BATTING_2B               -0.007472  0.009482 -0.788 0.430776  
## BATTING_3B                0.104289  0.017601  5.925 3.61e-09 ***
## BASERUN_SB                0.023849  0.005918  4.030 5.76e-05 ***
## BASERUN_CS                0.037198  0.016067  2.315 0.020690 *  
## PITCHING_H                0.008734  0.001205  7.248 5.80e-13 ***
## FIELDING_E                -0.035162  0.003271 -10.748 < 2e-16 ***
## FIELDING_DP               -0.084532  0.013704 -6.168 8.18e-10 ***
## ---                        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 12.74 on 2218 degrees of freedom
## Multiple R-squared:  0.2704, Adjusted R-squared:  0.2668
## F-statistic: 74.74 on 11 and 2218 DF,  p-value: < 2.2e-16

```



```

## Start: AIC=11321.3
## TARGET_WINS ~ BATTING_H + BATTING_2B + BATTING_3B + BATTING_HR +
##           BATTING_BB + BATTING_SO + BASERUN_SB + BASERUN_CS + PITCHING_H +
##           PITCHING_HR + PITCHING_BB + PITCHING_SO + FIELDING_E + FIELDING_DP +
##           BP_H + BP_HR + BP_BB + BP_SO
##
##
## Step: AIC=11321.3
## TARGET_WINS ~ BATTING_H + BATTING_2B + BATTING_3B + BATTING_HR +
##           BATTING_BB + BATTING_SO + BASERUN_SB + BASERUN_CS + PITCHING_H +
##           PITCHING_HR + PITCHING_BB + PITCHING_SO + FIELDING_E + FIELDING_DP +
##           BP_H + BP_HR + BP_BB
##
##
## Step: AIC=11321.3
## TARGET_WINS ~ BATTING_H + BATTING_2B + BATTING_3B + BATTING_HR +
##           BATTING_BB + BATTING_SO + BASERUN_SB + BASERUN_CS + PITCHING_H +
##           PITCHING_HR + PITCHING_BB + PITCHING_SO + FIELDING_E + FIELDING_DP +
##           BP_H + BP_HR
##
##
## Step: AIC=11321.3
## TARGET_WINS ~ BATTING_H + BATTING_2B + BATTING_3B + BATTING_HR +
##           BATTING_BB + BATTING_SO + BASERUN_SB + BASERUN_CS + PITCHING_H +
##           PITCHING_HR + PITCHING_BB + PITCHING_SO + FIELDING_E + FIELDING_DP +
##           BP_H
##
##
## Step: AIC=11321.3
## TARGET_WINS ~ BATTING_H + BATTING_2B + BATTING_3B + BATTING_HR +
##           BATTING_BB + BATTING_SO + BASERUN_SB + BASERUN_CS + PITCHING_H +
##           PITCHING_HR + PITCHING_BB + PITCHING_SO + FIELDING_E + FIELDING_DP
##
##          Df Sum of Sq   RSS   AIC
## - PITCHING_HR  1     71.3 352682 11320
## - BATTING_2B   1    147.1 352757 11320
## <none>                   352610 11321
## - BASERUN_CS   1    905.6 353516 11325
## - BATTING_HR   1    932.0 353542 11325
## - BATTING_H    1   1774.0 354384 11330
## - PITCHING_SO  1   1884.8 354495 11331
## - PITCHING_BB  1   2313.2 354924 11334

```

```

## - BATTING_SO  1  3164.7 355775 11339
## - BATTING_BB  1  3725.7 356336 11343
## - BASERUN_SB  1  4387.3 356998 11347
## - FIELDING_DP 1  5849.4 358460 11356
## - BATTING_3B  1  7439.0 360049 11366
## - PITCHING_H  1  12791.3 365402 11399
## - FIELDING_E  1  20259.4 372870 11444
##
## Step: AIC=11319.75
## TARGET_WINS ~ BATTING_H + BATTING_2B + BATTING_3B + BATTING_HR +
##             BATTING_BB + BATTING_SO + BASERUN_SB + BASERUN_CS + PITCHING_H +
##             PITCHING_BB + PITCHING_SO + FIELDING_E + FIELDING_DP
##
##          Df Sum of Sq   RSS   AIC
## - BATTING_2B  1    144.5 352826 11319
## <none>           352682 11320
## + PITCHING_HR 1     71.3 352610 11321
## + BP_HR        1     71.3 352610 11321
## - BASERUN_CS  1   923.1 353605 11324
## - BATTING_H   1   1786.1 354468 11329
## - PITCHING_SO 1   1855.1 354537 11329
## - BATTING_SO  1   3113.1 355795 11337
## - PITCHING_BB 1   4312.8 356995 11345
## - BASERUN_SB  1   4384.3 357066 11345
## - FIELDING_DP 1   5907.2 358589 11355
## - BATTING_BB  1   6344.8 359026 11358
## - BATTING_3B  1   7380.9 360063 11364
## - BATTING_HR  1  12111.1 364793 11393
## - PITCHING_H  1  12883.5 365565 11398
## - FIELDING_E  1  21170.8 373853 11448
##
## Step: AIC=11318.66
## TARGET_WINS ~ BATTING_H + BATTING_3B + BATTING_HR + BATTING_BB +
##             BATTING_SO + BASERUN_SB + BASERUN_CS + PITCHING_H + PITCHING_BB +
##             PITCHING_SO + FIELDING_E + FIELDING_DP
##
##          Df Sum of Sq   RSS   AIC
## <none>           352826 11319
## + BATTING_2B  1    144.5 352682 11320
## + PITCHING_HR 1     68.8 352757 11320
## + BP_HR        1     68.8 352757 11320
## - BASERUN_CS  1   997.3 353824 11323
## - PITCHING_SO 1   1748.3 354575 11328
## - BATTING_H   1   1953.4 354780 11329
## - BATTING_SO  1   3092.7 355919 11336
## - PITCHING_BB 1   4257.3 357084 11343
## - BASERUN_SB  1   4451.1 357277 11345
## - FIELDING_DP 1   5838.4 358665 11353
## - BATTING_BB  1   6265.6 359092 11356
## - BATTING_3B  1   7722.8 360549 11365
## - BATTING_HR  1  12449.8 365276 11394
## - PITCHING_H  1  13383.6 366210 11400
## - FIELDING_E  1  21030.9 373857 11446

```

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
	NA	NA	2215	352610.4	11321.30
- BP_SO	0	0.00000	2215	352610.4	11321.30
- BP_BB	0	0.00000	2215	352610.4	11321.30
- BP_HR	0	0.00000	2215	352610.4	11321.30
- BP_H	0	0.00000	2215	352610.4	11321.30
- PITCHING_HR	1	71.32528	2216	352681.7	11319.75
- BATTING_2B	1	144.53570	2217	352826.2	11318.66

```

## 
## Call:
## lm(formula = TARGET_WINS ~ BATTING_H + BATTING_3B + BATTING_HR +
##      BATTING_BB + BATTING_SO + BASERUN_SB + BASERUN_CS + PITCHING_H +
##      PITCHING_BB + PITCHING_SO + FIELDING_E + FIELDING_DP, data = imputed_train)
## 
## Residuals:
##    Min      1Q  Median      3Q     Max 
## -80.887 -8.130   0.041   8.097  68.359 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 33.175729  5.656513  5.865 5.16e-09 ***
## BATTING_H    0.013834  0.003949  3.503 0.000468 ***
## BATTING_3B   0.125448  0.018008  6.966 4.28e-12 ***
## BATTING_HR   0.090183  0.010196  8.845 < 2e-16 ***
## BATTING_BB   0.083849  0.013363  6.275 4.20e-10 ***
## BATTING_SO   -0.029123 0.006606 -4.408 1.09e-05 ***
## BASERUN_SB   0.032719  0.006187  5.289 1.35e-07 ***
## BASERUN_CS   0.040007  0.015982  2.503 0.012375 *  
## PITCHING_H   0.014048  0.001532  9.170 < 2e-16 ***
## PITCHING_BB  -0.060007 0.011602 -5.172 2.52e-07 ***
## PITCHING_SO  0.017834  0.005381  3.314 0.000933 *** 
## FIELDING_E   -0.037924 0.003299 -11.496 < 2e-16 ***
## FIELDING_DP  -0.082243 0.013579 -6.057 1.63e-09 ***
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 12.62 on 2217 degrees of freedom
## Multiple R-squared:  0.2849, Adjusted R-squared:  0.2811 
## F-statistic: 73.62 on 12 and 2217 DF,  p-value: < 2.2e-16
## 
## Call:
## lm(formula = TARGET_WINS ~ BATTING_3B + BATTING_HR + BATTING_BB +
##      PITCHING_BB + BATTING_SO + BASERUN_SB + BASERUN_CS + log(PITCHING_H) +
##      log(FIELDING_E) + FIELDING_DP, data = imputed_train)
## 
## Residuals:
##    Min      1Q  Median      3Q     Max 
## -65.732 -7.705   0.020   7.770  70.422 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -1.565e+02  2.608e+01 -6.001 2.28e-09 ***
## BATTING_3B   1.598e-01  1.635e-02  9.771 < 2e-16 ***

```

```

## BATTING_HR          6.900e-02  9.843e-03   7.010 3.14e-12 ***
## BATTING_BB          7.055e-02  1.446e-02   4.880 1.14e-06 ***
## PITCHING_BB         -7.169e-02 9.991e-03  -7.175 9.79e-13 ***
## BATTING_SO          -1.133e-02 2.130e-03  -5.319 1.15e-07 ***
## BASERUN_SB           2.978e-02  5.672e-03   5.251 1.66e-07 ***
## BASERUN_CS           5.396e-02  1.574e-02   3.427 0.000621 ***
## log(PITCHING_H)     4.287e+01  3.172e+00  13.517 < 2e-16 ***
## log(FIELDING_E)     -1.604e+01 1.062e+00 -15.101 < 2e-16 ***
## FIELDING_DP          -9.269e-02 1.322e-02  -7.012 3.11e-12 ***
## BATTING_BB:PITCHING_BB 2.575e-05  1.544e-05   1.668 0.095489 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.4 on 2218 degrees of freedom
## Multiple R-squared:  0.3084, Adjusted R-squared:  0.3049
## F-statistic:  89.9 on 11 and 2218 DF,  p-value: < 2.2e-16

```

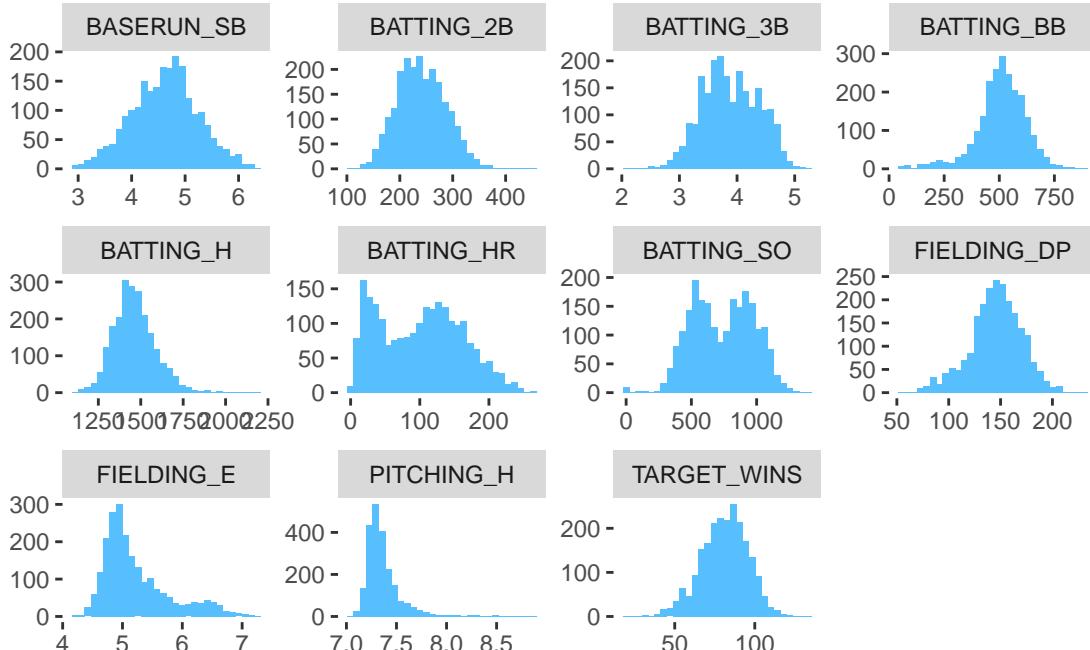
---

### 3.4 MODEL 3

We sought to explore whether there was a relationship between wins and the difference of specific offensive and defensive team capabilities - hits, homeruns, balls, and strike-outs. Incorporating variables that reflect those differences (i.e. subtracting batting hits from pitching hits, and so on), however, did not improve the explanatory power of the model beyond using the original variables.

Given these variables did not yield improvements, in their place we explored a third model. As the histograms below highlight, a number of the independent variables - pitching hits, pitching homeruns, pitching strikeouts - demonstrate pronounced rightward-skew.

Histograms



We corrected for that skew by transforming those three variables using natural logarithms. When we tested those log transformations in a model where they replaced the untransformed, original variables combined

with all other variables, we found that neither the originals nor the log transformations for pitching homeruns and pitching strikeouts met the threshold of significance (a p-value below the  $\alpha$  level of .05). Based on high p-values, over a series of backward steps we removed pitching homeruns, pitching strikeouts, and baserun caught stealing, yielding the following model:

[Jeremy: team, should we write LaTeX formulas for each model or just cable model coefficients?]

```
$y = $  
##  
## Call:  
## lm(formula = TARGET_WINS ~ BATTING_H + BATTING_2B + BATTING_3B +  
##      BATTING_HR + BATTING_BB + BATTING_SO + BASERUN_SB + log(PITCHING_H) +  
##      PITCHING_BB + FIELDING_E + FIELDING_DP, data = imputed_train)  
##  
## Residuals:  
##      Min      1Q Median      3Q     Max  
## -78.012 -8.149   0.253   7.985  73.220  
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -4.050e+02  3.700e+01 -10.947 < 2e-16 ***  
## BATTING_H    -1.346e-02  6.034e-03 -2.232  0.02574 *  
## BATTING_2B   -1.500e-02  9.146e-03 -1.641  0.10102  
## BATTING_3B    1.405e-01  1.766e-02  7.960 2.72e-15 ***  
## BATTING_HR   7.230e-02  9.704e-03  7.451 1.32e-13 ***  
## BATTING_BB   1.247e-01  1.266e-02  9.846 < 2e-16 ***  
## BATTING_SO   -6.554e-03  2.330e-03 -2.813  0.00495 **  
## BASERUN_SB   4.380e-02  4.763e-03  9.197 < 2e-16 ***  
## log(PITCHING_H) 6.866e+01  5.760e+00 11.921 < 2e-16 ***  
## PITCHING_BB  -9.518e-02  1.063e-02 -8.953 < 2e-16 ***  
## FIELDING_E   -4.740e-02  3.550e-03 -13.350 < 2e-16 ***  
## FIELDING_DP  -8.853e-02  1.294e-02 -6.842 1.01e-11 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 12.49 on 2218 degrees of freedom  
## Multiple R-squared:  0.2991, Adjusted R-squared:  0.2956  
## F-statistic: 86.03 on 11 and 2218 DF,  p-value: < 2.2e-16
```

Based on this model's F-statistic and p-value, we can reject the null hypothesis that coefficients with values of zero would fit the data better. Per the adjusted  $r^2$  value, this model explains approximately 29.56% of the variance in wins. However, in doing so it treats the batting hits and batting second base runs as drags on wins (with negative coefficients), and pitching hits as buoying wins - which is counterintuitive. While the other coefficients make more intuitive sense, these signs call into question how effectively we can use this model to understand the relationships between the independent variables and wins.

## 4 SELECT MODELS

### 4.1 Instructions:

Decide on the criteria for selecting the best multiple linear regression model. Will you select a model with slightly worse performance if it makes more sense or is more parsimonious? Discuss why you selected your

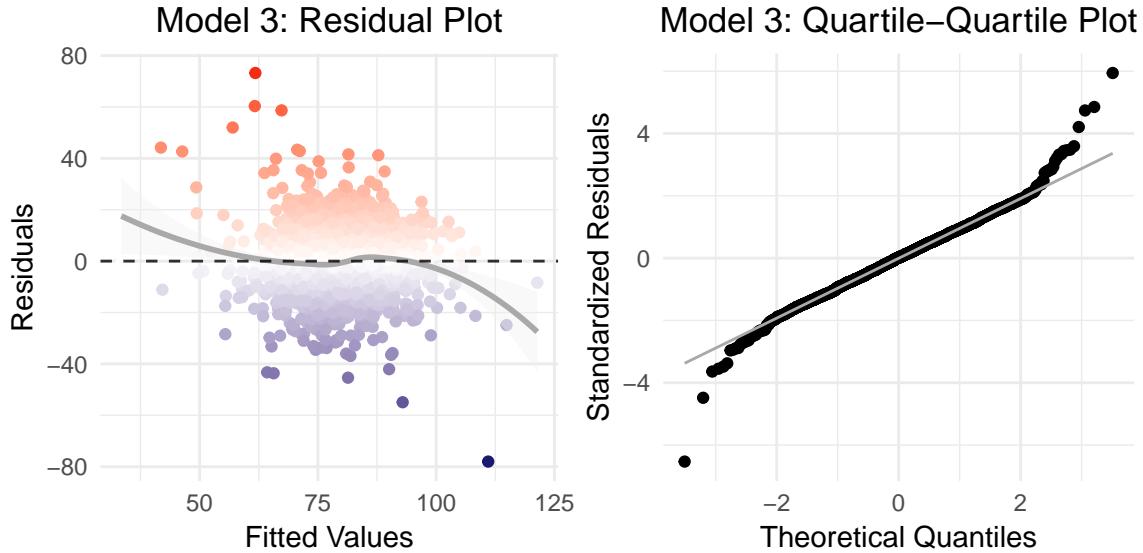
Table 3: Log Transform Model Coefficients

	x
(Intercept)	-405.0139425
BATTING_H	-0.0134654
BATTING_2B	-0.0150055
BATTING_3B	0.1405383
BATTING_HR	0.0723000
BATTING_BB	0.1246742
BATTING_SO	-0.0065541
BASERUN_SB	0.0438045
log(PITCHING_H)	68.6642334
PITCHING_BB	-0.0951797
FIELDING_E	-0.0473986
FIELDING_DP	-0.0885260

model. For the multiple linear regression model, will you use a metric such as Adjusted R<sup>2</sup>, RMSE, etc.? Be sure to explain how you can make inferences from the model, discuss multi-collinearity issues (if any), and discuss other relevant model output. Using the training data set, evaluate the multiple linear regression model based on (a) mean squared error, (b) R<sup>2</sup>, (c) F-statistic, and (d) residual plots. Make predictions using the evaluation data set.

## 4.2 Comparison of models

[Jeremy: added in a chart for model 3]



## 5 Multi-collinearity

[Jeremy: We can place the correlogram here or revisit it]

An examination of the partial correlation coefficients between all the independent variables shows that there are the expected relationships between the offense and defense variables - i.e. batting and pitching homeruns -

are there, meet p-value thresholds, and are strong. This finding suggests keeping only one of each pair of variables in a given model; yet, when we tried this, less of the variance in wins was explained by the model.

	BATTING_2B	BATTING_3B	BATTING_HR	BATTING_BB	BATTING_SO	BASERUN
BATTING_2B	1.0000000	0.2129669	0.0609690	0.1206106	-0.0764713	0.107
BATTING_3B	0.2129669	1.0000000	-0.0730874	0.0131362	-0.1506443	0.171
BATTING_HR	0.0609690	-0.0730874	1.0000000	0.5747337	0.1124715	0.015
BATTING_BB	0.1206106	0.0131362	0.5747337	1.0000000	0.5064157	-0.160
BATTING_SO	-0.0764713	-0.1506443	0.1124715	0.5064157	1.0000000	0.337
BASERUN_SB	0.1073938	0.1713437	0.0156533	-0.1602112	0.3375119	1.000
BASERUN_CS	-0.1126072	0.0577696	-0.0273953	-0.0008126	-0.1012046	0.617
PITCHING_H	0.3500087	-0.1142022	0.1033127	-0.3937551	-0.0524595	-0.079
PITCHING_HR	-0.0287742	0.0540702	0.9817947	-0.5681054	-0.0702943	-0.000
PITCHING_BB	-0.1060008	-0.0071098	-0.5899436	0.9778943	-0.5319800	0.225
PITCHING_SO	0.0983961	0.0219166	-0.0598031	-0.5935317	0.9486979	-0.308
FIELDING_E	-0.2553727	0.1729403	0.2509843	-0.0913065	-0.2047899	0.220
FIELDING_DP	0.0551834	0.0042580	-0.0212282	0.0455742	-0.0470879	-0.013
	BATTING_2B	BATTING_3B	BATTING_HR	BATTING_BB	BATTING_SO	BASERUN
BATTING_2B	0.0000000	0.0000000	0.0040650	0.0000000	0.0003115	0.000
BATTING_3B	0.0000000	0.0000000	0.0005699	0.5362626	0.0000000	0.000
BATTING_HR	0.0040650	0.0005699	0.0000000	0.0000000	0.0000001	0.461
BATTING_BB	0.0000000	0.5362626	0.0000000	0.0000000	0.0000000	0.000
BATTING_SO	0.0003115	0.0000000	0.0000001	0.0000000	0.0000000	0.000
BASERUN_SB	0.0000004	0.0000000	0.4611241	0.0000000	0.0000000	0.000
BASERUN_CS	0.0000001	0.0064878	0.1970481	0.9694840	0.0000018	0.000
PITCHING_H	0.0000000	0.0000001	0.0000011	0.0000000	0.0134555	0.000
PITCHING_HR	0.1754304	0.0108506	0.0000000	0.0000000	0.0009212	0.978
PITCHING_BB	0.0000006	0.7378265	0.0000000	0.0000000	0.0000000	0.000
PITCHING_SO	0.0000034	0.3020954	0.0048318	0.0000000	0.0000000	0.000
FIELDING_E	0.0000000	0.0000000	0.0000000	0.0000165	0.0000000	0.000
FIELDING_DP	0.0093223	0.8411154	0.3175375	0.0318142	0.0265469	0.533
	BATTING_2B	BATTING_3B	BATTING_HR	BATTING_BB	BATTING_SO	BASERUN
BATTING_2B	0.000000	10.2629918	2.8760766	5.7207133	-3.611228	5.086
BATTING_3B	10.262992	0.0000000	-3.4505512	0.6185709	-7.174974	8.188
BATTING_HR	2.876077	-3.4505512	0.0000000	33.0685744	5.329542	0.737
BATTING_BB	5.720713	0.6185709	33.0685744	0.0000000	27.652657	-7.642
BATTING_SO	-3.611228	-7.1749738	5.3295425	27.6526571	0.000000	16.882
BASERUN_SB	5.086057	8.1888267	0.7371254	-7.6422669	16.882395	0.000
BASERUN_CS	-5.336053	2.7246343	-1.2903952	-0.0382598	-4.789815	36.969
PITCHING_H	17.592987	-5.4126239	4.8906535	-20.1693441	-2.473462	-3.752
PITCHING_HR	-1.355393	2.5496259	243.3742845	-32.5038836	-3.318017	-0.027
PITCHING_BB	-5.019328	-0.3347749	-34.4017617	220.2020200	-29.581449	10.907
PITCHING_SO	4.655576	1.0321907	-2.8208809	-34.7242694	141.276857	-15.288
FIELDING_E	-12.436593	8.2674688	12.2083774	-4.3172042	-9.851330	10.665
FIELDING_DP	2.602278	0.2004905	-0.9997564	2.1480957	-2.219598	-0.622